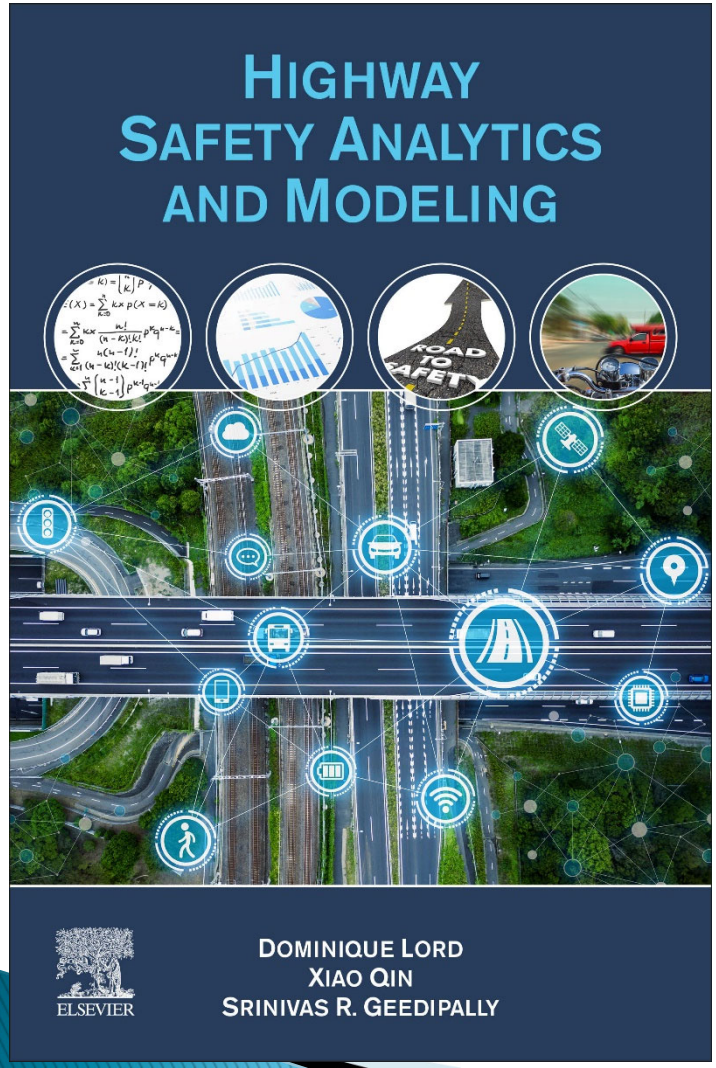
Crash-Frequency Models

Part 1

November 3, 2021

Instructor: Dominique Lord
Texas A&M University
d-lord@tamu.edu

Textbook



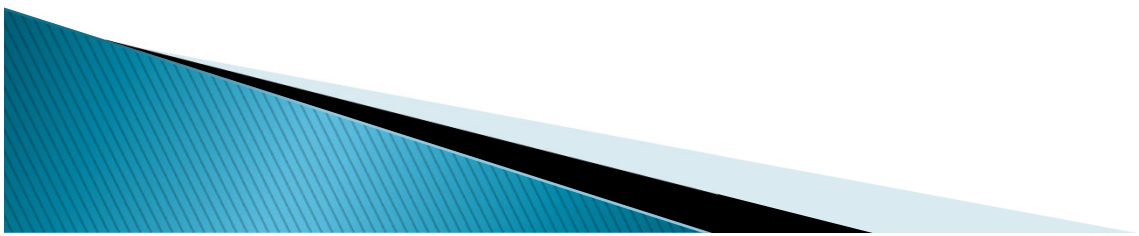
The material presented in this series of lectures are taken from this textbook and other sources based on lectures given by the authors.

The textbook is available on Amazon and the Elsevier website below among other places.

https://www.elsevier.com/books/highway-safety-analytics-and-modeling/lord/978-0-12-816818-9

Why use Statistical Models?

- Crashes are “independent” and “random” events (probabilistic events)
- Estimate a relationship between crashes and covariates (or explanatory variables)
- Determine the long-term average of crash occurrences for transportation facilities
- Have a wide variation of applications in safety analyses:
  - Prediction
  - Variable screening
  - Risk factors
  - Before-after study



Statistical Models For Crash Data

Modeling Process

1. Determine Modeling Objectives

- Definition (Intersections, Pedestrians, etc.)
- Data availability
- Unit Scales (Crashes/year; Severity; etc.)

2. Establish Appropriate Process

- Sampling Models
- Observational Models
- Process/System State Models
- Parameter Models (Bayesian Models Only)

Statistical Models For Crash Data

Modeling Process

3. Determine Inferential Goals

- Point estimate (Value + Standard Error)
- Distribution (Bayesian Models)
- Percentiles (2.5%, 85%, etc.; Bayesian Models)

↓

4. Select Computation Techniques

- Frequentist (MLE)
- Bayesian (via simulation)
- Empirical Bayes

↓

5. Evaluate Models

- Goodness-of-Fit
- Prediction
- Confidence Intervals

Basic Nomenclature

$$y_i = f\left(\mathbf{x}'\boldsymbol{\beta}_i\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

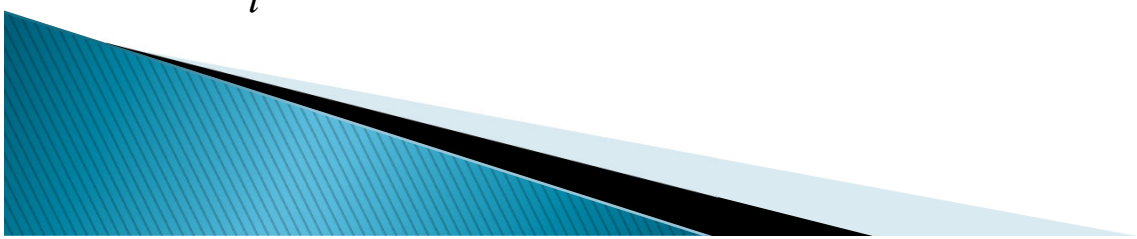
y_i is the response variable for observation *i.*

$\boldsymbol{\beta}_i$ is a p x 1 vector of estimable parameters.

$\mathbf{x}'$ is a vector of explanatory variables.

p is the number of parameters in the model.

ε_i is a random error term of the model.



Basic Nomenclature

The previous equation can be re-written as follows

$$E[y_i|\mathbf{x}_i] = \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

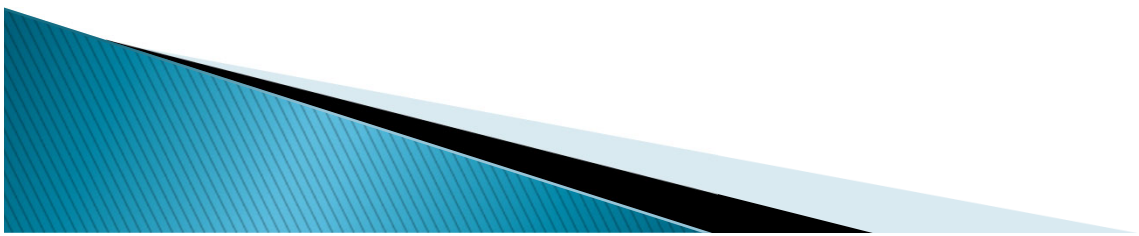
Based on the generalized linear modeling relationship with an exponential canonical link function, the equation leads to the following form:

$$\mu_i = \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) = \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right).$$

All the models described further below will follow the form described above.

Sources of Dispersion

- Over-Dispersion
  - Unequal probability of events (Poisson trials)
  - Unobserved heterogeneity (crash rate that differs across observations)
    - Factors that influence crash risk not captured by the data/model
- Under-Dispersion (rare)
  - Two Conditions
  - 1) Low sample mean
  - 2) Modeling output (Observations conditional upon the mean)



Sources of Dispersion

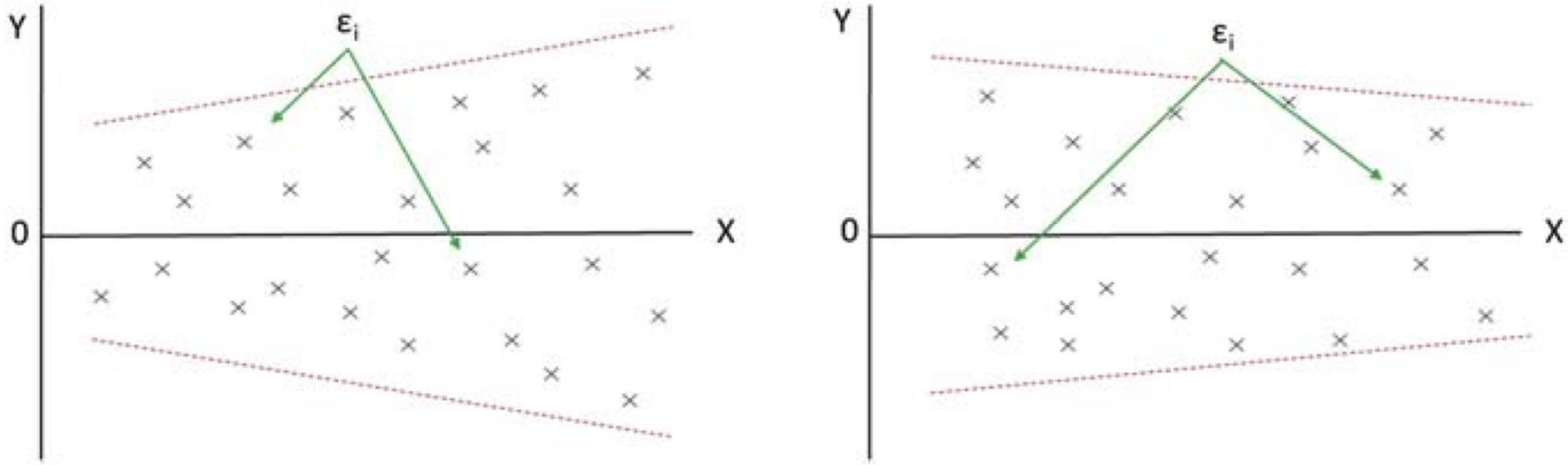


FIGURE 3.1 Overdispersed (left) and underdispersed (right) residuals.

$$Var[y_i] > E[y_i]$$

$$Var[y_i] < E[y_i]$$

Basic Count Models

Poisson Model

In a Poisson regression model, the probability of a roadway entity (segment, intersection, vehicle, etc.) i having y_i crashes per some time period (where y_i is a non-negative integer) is given by:

$$P(y_i \mid \mathbf{x}_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$$

$P(y_i \mid \mathbf{x}_i)$ is the probability of roadway entity (or observation) *i* having y_i crashes per time period.

$\mu_i = \exp(\mathbf{x}_i' \boldsymbol{\beta})$ is the Poisson mean parameter for roadway entity *i*.

$Var[y_i] = E[y_i]$ is extremely rare.

Basic Count Models

Poisson-gamma Model (NB)

The PMF of the Poisson-gamma regression for y_i is

$$P(y_i \mid \mathbf{x}_i, \alpha) = \frac{\Gamma(1/\alpha + y_i)}{\Gamma(1/\alpha) y_i!} \left(\frac{1/\alpha}{(1/\alpha) + \mu_i} \right)^{1/\alpha} \left(\frac{\mu_i}{(1/\alpha) + \mu_i} \right)^{y_i}$$

The mean and variance are given by

$$E(y_i \mid \mathbf{x}_i) = u_i$$

$$Var(y_i \mid \mathbf{x}_i) = u_i + \alpha\mu_i^2 \text{ or } Var(y_i \mid \mathbf{x}_i) = u_i + \frac{\mu_i^2}{\phi}$$

The mean function is given by

$$E(y_i \mid \mathbf{x}_i) = u_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i) \quad \exp(\varepsilon_i) \sim \text{gamma}(1, 1/\phi)$$

Basic Count Models

Poisson-gamma Model

Example – Crash Data at 3-legged signalized intersections:

Functional form: $\mu = e^{\beta_0 + \beta_1 F_{maj} + \beta_2 F_{maj}}$

Functional form needed to model crash data:

$$\mu = \beta_0 F_{maj}^{\beta_1} F_{\min}^{\beta_2}$$

Need to take the natural log of the flow variables

Where,

$\mu =$ Expected number of crashes

$F_{maj} =$ Major traffic flow

$F_{\min} =$ Minor traffic flow

Basic Count Models

Poisson-gamma Model

$$\mu = e^{-10.065} F_{maj}^{0.751} F_{\min}^{0.484}$$

$$\mu = 4.05E-05 \times F_{maj}^{0.751} F_{\min}^{0.484}$$

$$Var(y) = \mu + 0.315\mu^2$$

The GENMOD Procedure

Model Information

Data Set	WORK.C
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	Total Total

Number of Observations Read	255
Number of Observations Used	255

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	252	288.8580	1.1463
Scaled Deviance	252	288.8580	1.1463
Pearson Chi-Square	252	312.6975	1.2409
Scaled Pearson X2	252	312.6975	1.2409
Log Likelihood		836.0686	
Full Log Likelihood		-606.7989	
AIC (smaller is better)		1221.5978	
AICC (smaller is better)		1221.7578	
BIC (smaller is better)		1235.7628	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-10.0648	1.3659	-12.7420	-7.3876	54.29	<.0001
logf_maj	1	0.7517	0.1320	0.4929	1.0105	32.41	<.0001
logf_min	1	0.4837	0.0562	0.3735	0.5939	74.01	<.0001
Dispersion	1	0.3153	0.0519	0.2135	0.4170		

NOTE: The negative binomial dispersion parameter was estimated by maximum likelihood.

Basic Count Models

Poisson-Lognormal Model (PLN)

The PMF of the Poisson-lognormal regression is not available for the PLN, since it does not have a closed form.

The mean and variance are given by

$$E\left[y_i \mid \mathbf{x}_i\right] = \mu_i = \exp\left(\mathbf{x}_i'\boldsymbol{\beta} + 1/\sigma^2 + \varepsilon_i\right)$$

$$Var(y_i \mid \mathbf{x}_i) = e^{\mu_i + \sigma^2/2} + \left[e^{\sigma^2} - 1\right] e^{2\mu_i + \sigma^2}$$

The error is given by

$$\varepsilon_i \sim \text{Normal}\left(0, \sigma^2\right) \quad \text{Note:} \quad \exp\left(\varepsilon_i\right) \sim \text{Lognormal}\left(0, \sigma^2\right)$$

Basic Count Models

Poisson-Lognormal Model (PLN)

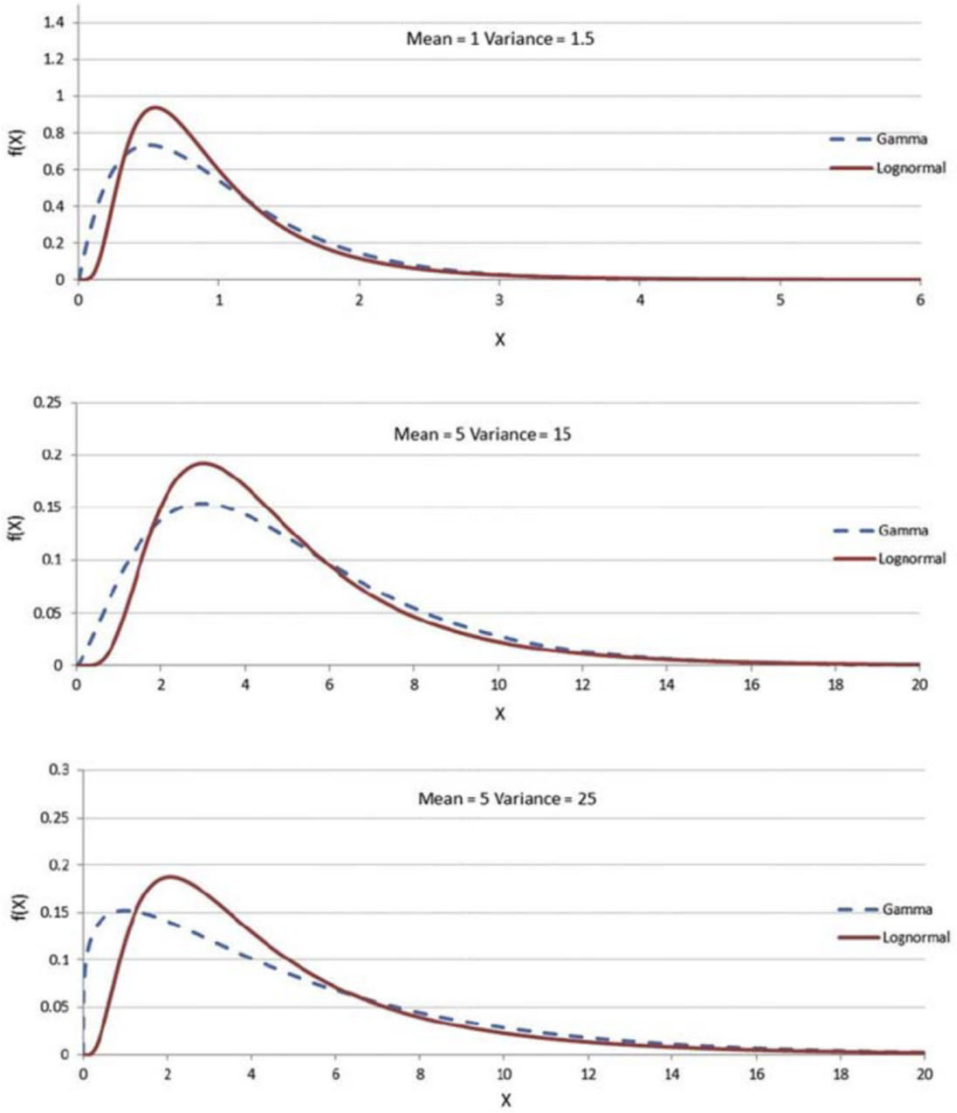
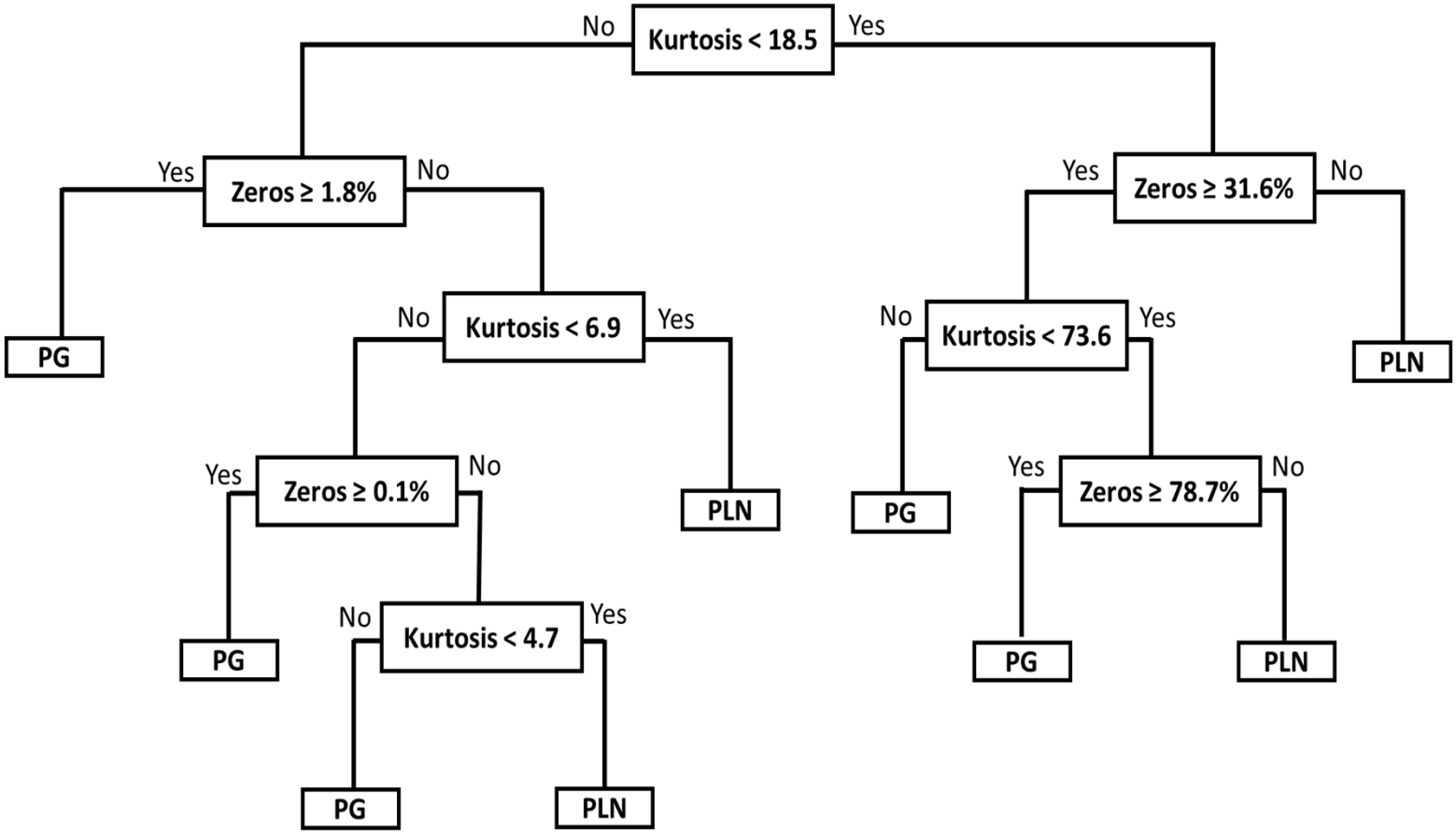


FIGURE 3.2 Probability distribution function of the Poisson-gamma (NB) and lognormal distributions for different mean-variance combinations (Khazraee et al., 2018).

Basic Count Models

Selecting between PG/NB and PLN



Calculate the % of zeros and Kurtosis with the data at hand. Then, follow the tree branches above.

Basic Count Models

Some of these models include:

Poisson-Weibull: The Poisson-Weibull distribution model performs as well as the NB model and its coefficients can be easily estimated using the MLE.

Poisson-Inverse Gaussian (PIG): The PIG model performs similar to the PLN, in that the model fits the data better at the tail end of the distribution. The coefficients are also easily estimated using the MLE.

Poisson-Inverse Gamma: This model also performs similar to the PLN for long-tailed data. The model is estimated using the Bayesian estimating method, which requires further work than the MLE.

Sichel [SI]: This model has been used or applied more frequently than the previous models. The SI model is recommended to be used for long-tailed data. This model can be estimated using the MLE.

Poisson-Tweedie: Depending on the parameterization of the model, the Poisson-Tweedie distribution models can become as special cases to the NB, PIG, or SI model. The same characteristics as those listed above apply here.

Generalized count models for underdispersion

Conway-Maxwell-Poisson (COM-Poisson)

The PMF of the COM-Poisson regression for y_i is

$$P\left(y_i \mid \mathbf{x}_i\right) = \frac{1}{Z\left(\mu_i, \nu\right)}\left(\frac{{\mu_i}^{y_i}}{y_i!}\right)^{\nu} \qquad Z\left(\mu_i, \nu\right) = \sum_{n=0}^{\infty}\left(\frac{{\mu_i}^{n}}{n!}\right)^{\nu}$$

The mean and variance are given by

$$\mu_i = \exp\left(\beta_0 + \sum_{i=1}^{p}\beta_i x_i\right)$$

$$\nu_i = \exp\left(\gamma_0 + \sum_{j=1}^{q}\gamma_j x_j\right)$$

ν > 1 -> Over-Dispersion

ν < 1 -> Under-Dispersion

ν = 1 -> Poisson

Note: $\mu_i = \lambda_i^{1/\nu}$.

Generalized count models for underdispersion

Other Models

Other models that have been proposed for analyzing underdispersion include the following:

Gamma model (continuous distribution): This model cannot account for observations with zero counts. It is presented here as a caution for not using this model.

Gamma-count model: This modified gamma model has been proposed by Winkelman (1995). The parameterization offered by this researcher assumes that the observations have a direct correlation with each other in time. In safety, this means that a crash at time t is directly related to a crash at time t + n, which is again theoretically impossible.

Double Poisson: This model has initially been proposed by Efron (1986). However, it has not been used often, as the normalizing constant of the model is not properly definite. Zou et al. (2013) have proposed a different parameterization of the constant term and found results similar to the COM-Poisson.

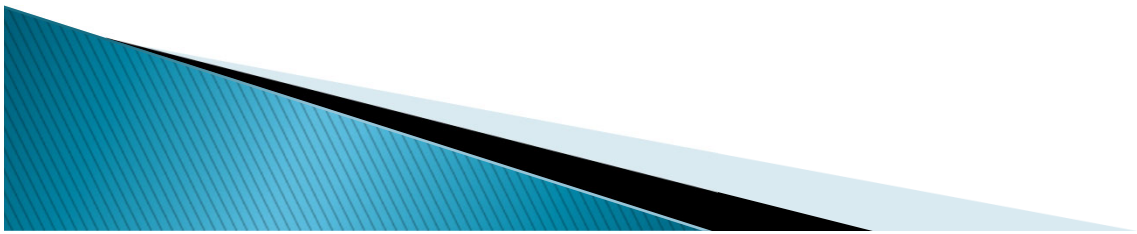
Generalized count models for underdispersion

Other Models

Other models that have been proposed for analyzing underdispersion include the following:

Hyper-Poisson: The hyper-Poisson (hP) is a two-parameter generalization of the Poisson distribution. Similar to the COM-Poisson, it can model the variance function as a function of the covariates. It performs as well as the COM-Poisson and can be estimated using the MLE.

Generalized Event Count: This model uses the theoretical statistics called "bilinear recurrence relationship" that was introduced by Katz (1965) for describing the dispersion parameter of the Poisson count model. Ye et al. (2018) applied the model to crash data and found its performance to be similar to the hP.



Finite mixture and multivariate models

Finite Mixture Model - Poisson-Gamma Model (NB)

The PMF of the FMNB-K regression for y_i is

$$P\left(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}\right) = \sum_{k=1}^{K} w_k \mathrm{NB}\left(\mu_{k,i}, \phi_k\right) = \sum_{k=1}^{K} w_k \left[\frac{\Gamma\left(y_i + \phi_k\right)}{\Gamma\left(y_i + 1\right)\Gamma\left(\phi_k\right)} \left(\frac{\mu_{k,i}}{\mu_{k,i} + \phi_k} \right)^{y_i} \left(\frac{\phi_k}{\mu_{k,i} + \phi_k} \right)^{\phi_k} \right]$$

The mean and variance are given by

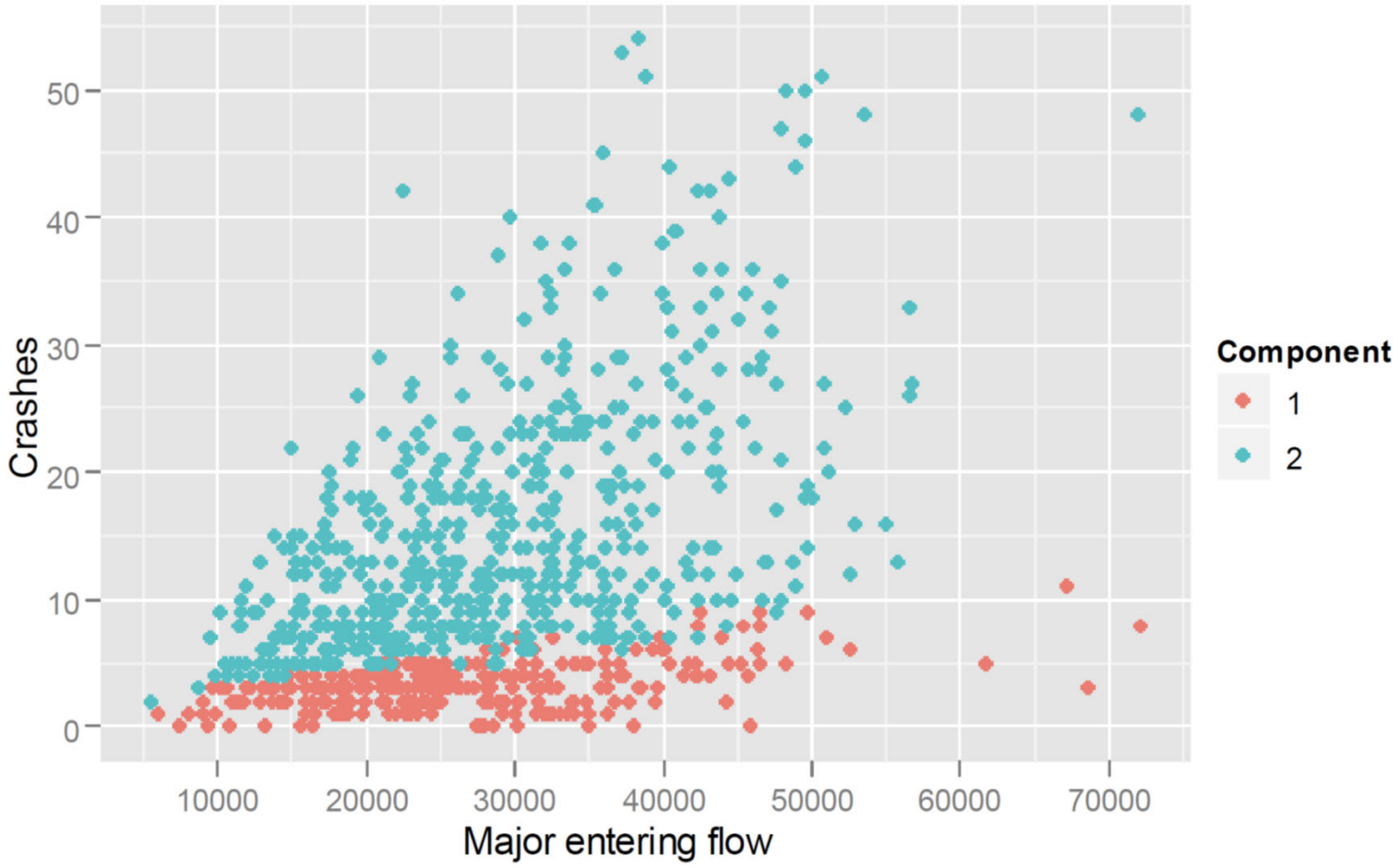
$$E\left(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}\right) = \sum_{k=1}^{K} w_k \mu_{k,i}$$

$$Var\left(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}\right) = E\left(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}\right) + \left(\sum_{k=1}^{K} w_k \mu_{k,i}^2 \left(1 + {}^{1}\!/_{\phi_k} \right) - E\left(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}\right)^2 \right)$$

Note: Each class will have their own coefficients or parameters.

Finite mixture and multivariate models

Finite Mixture Model - Poisson-Gamma Model (NB)



(a) The number of crashes against the major entering flow

Finite mixture and multivariate models

Multivariate models

Each parameter and explanatory variables are vectors/matrices:

$$\mathbf{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1m} \\ \vdots & \ddots & \vdots \\ \beta_{p1} & \cdots & \beta_{pm} \end{bmatrix}$$

$$P(y_{im}|\mathbf{x}_i, b_i, \beta_m) \tilde{} \text{Poisson}(\mu_{im})$$

$$\mu_{im} = \exp(x_i \beta_m + b_{im})$$

These models are used to account for the correlation between crash severity levels or collision types. The most common model is the multivariate Poisson-lognormal model. The multivariate NB has been proposed but suffers from important methodological limitations.

Models for better capturing unobserved heterogeneity

Random-effects/multilevel model

Random-effects (RE) models, or sometimes called multilevel models, are models that allow the variance that may exist within different levels of the data to be better depicted. This is accomplished by adding one or more RE terms or random intercept term to capture the between observations variance. Taking the basic models described above, the formulation becomes

$$\mu_{io} = \exp\left(\mathbf{x}'_{io}\boldsymbol{\beta} + \varpi_o + \varepsilon_{io}\right)$$

Models for better capturing unobserved heterogeneity

Random-Parameters NB Models (RPNB)

Same PMF as before for y_i is

$$P(y_i \mid \mathbf{x}_i, \alpha) = \frac{\Gamma(1/\alpha + y_i)}{\Gamma(1/\alpha) y_i!} \left(\frac{1/\alpha}{(1/\alpha) + \mu_i} \right)^{1/\alpha} \left(\frac{\mu_i}{(1/\alpha) + \mu_i} \right)^{y_i}$$

The parameters can have a mean and variance:

$$\mu_i = \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) = \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i\right)$$

$$\beta_{ik} = b_k + \upsilon_{ik} \quad \text{or} \quad \beta_i = \mathbf{b} + \boldsymbol{\Theta}\mathbf{z}_i$$

There is also a version that account for correlation between the variables.

Models for better capturing unobserved heterogeneity

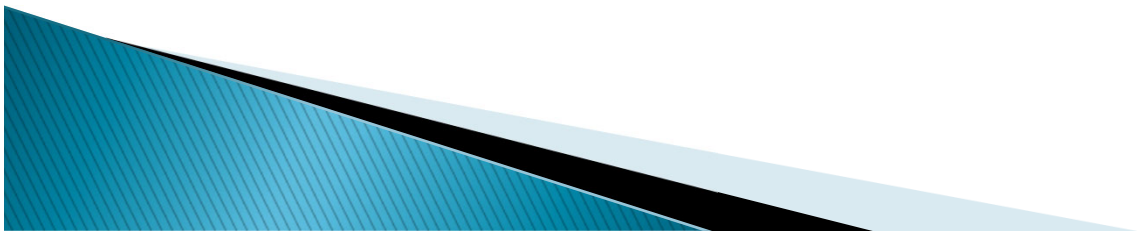
Random-Parameters NB Models (RPNB)

The log-likelihood is defined like this

$$LL = \sum_{\forall i} \ln \int_{\varphi_i} g(\varphi_i) P(y_i|\varphi_i)\, d\varphi_i$$

where g (φ_i) is the probability density function of the φ_i.

The log-likelihood needs to be estimated using simulation, such as the Monte Carlo simulation (not MCMC).



Models for better capturing unobserved heterogeneity
Random-Parameters NB Models (RPNB)

- Although RP models reduce the unobserved heterogeneity, some of the coefficients may be difficult to explain or describe.
- Usually, only some parameters are random, while others are fixed.
- The issue related to Bayesian estimation vs RP models still need to be fully addressed
  - Under the Bayesian estimation approach, all the parameters are considered random. Hence, why not use Bayesian and assume all variables to be random?
- Should not only be used for improving the GOF

Models for better capturing unobserved heterogeneity

Random-Parameters NB Models (RPNB)

Table 2
Parameter estimates from different random parameters types of negative binomial count models.

Variables	Uncorrelated random parameters model		Correlated random parameters model	
	Estimates	t-Value	Estimates	t-Value
Gross value added	−0.104***	−31.19	−0.131***	−41.44
Employment rate	0.005***	5.69	0.005***	5.55
Urbanization rate	0.034***	62.51	0.018***	35.43
Education share	−0.008***	−4.77	−0.005***	−3.34
Health personnel share	0.581***	26.16	1.158***	19.45
Hospital capacity share	−0.286	−6.43	−0.268	−5.75
Transportation share	−0.049***	−14.40	0.035***	11.76
Alcohol share	−0.224***	−20.99	−0.260***	−25.97
Migration dummy	−0.060***	−5.13	−0.039***	−3.49
Highways	−0.358***	−28.19	0.670***	52.48
Marmara region	0.813***	33.66	−0.109***	−4.53
Aegean region	1.545***	52.40	0.261***	10.61
Mediterranean region	1.656***	50.48	−0.373***	−14.16
Central Anatolia region	1.327***	55.51	0.007	0.31
Black Sea region	0.750***	38.31	−0.030	−1.58
Southeastern Anatolia	0.496***	20.93	0.178***	7.80
Year 2009	0.052***	4.95	0.069***	6.62
Year 2010	0.109***	6.98	0.144***	10.09
Winter	−0.257***	−16.13	−0.254***	−15.82
Spring	0.048***	3.00	0.051***	3.18
Summer	0.198***	10.84	0.203***	10.78
Migration dummy*Summer	0.106***	10.03	0.103***	9.66
Migration dummy*Fall	0.119***	6.71	0.121***	6.85
Means for random parameters:				
Constant	1.871***	20.95	2.701***	31.08
Vehicles share	0.029***	22.00	0.026***	20.55
Rain amount	−0.001***	−7.33	−0.001***	−6.53
Share of the red light rule violation	0.002*	1.74	0.002***	2.83
Share of the speed limit rule violation	0.004***	5.75	0.004***	11.00
Scale parameters for distribution of random parameters:				
Constant	0.485***	71.11	0.650***	28.55
Vehicles share	0.076***	84.25	0.001	1.53
Rain amount	0.001***	13.56	0.000***	5.58
Share of the red light rule violation	0.011***	22.02	0.009***	14.30
Share of the speed limit rule violation	0.040***	33.49	0.003***	24.65
Dispersion parameter (α)	69.432***	22.41	67.814***	22.57
Other statistical information:				
Log-likelihood values at convergence (LL_c)		−11,247.824		−11,217.075
Restricted log-likelihood value (LL_r)		−1,179,724.853		−1,179,724.853
AIC		22,563.648		22,522.150
BIC		22,766.899		22,785.181
Pseudo-R^2		0.304		0.306
Predicted numbers of accidents		94.180		94.161

Note that the degrees of freedom for LL_r is 5 and 15 for uncorrelated and correlated random parameters models since the model compare them to the conventional negative binomial model with pooled data.
* $p < 0.10$;** $p < 0.05$;*** $p < 0.01$.

Models for better capturing unobserved heterogeneity

Random-Parameters NB Models (RPNB)

Table 2
Parameter estimates from different random parameters types of negative binomial count models.

Variables	Uncorrelated random parameters model		Correlated random parameters model	
	Estimates	*t*-Value	Estimates	*t*-Value
Gross value added	−0.104***	−31.19	−0.131***	−41.44
Employment rate	0.005***	5.69	0.005***	5.55
Urbanization rate	0.034***	62.51	0.018***	35.43
Education share	−0.008***	−4.77	−0.005***	−3.34
Health personnel share	0.581***	26.16	1.158***	19.45
Hospital capacity share	−0.286	−6.43	−0.268	−5.75
Transportation share	−0.049***	−14.40	0.035***	11.76
Alcohol share	−0.224***	−20.99	−0.260***	−25.97
Migration dummy	−0.060***	−5.13	−0.039***	−3.49
Highways	−0.358***	−28.19	0.670***	52.48
Marmara region	0.813***	33.66	−0.109***	−4.53
Aegean region	1.545***	52.40	0.261***	10.61
Mediterranean region	1.656***	50.48	−0.373***	−14.16
Central Anatolia region	1.327***	55.51	0.007	0.31
Black Sea region	0.750***	38.31	−0.030	−1.58
Southeastern Anatolia	0.496***	20.93	0.178***	7.80
Year 2009	0.052***	4.95	0.069***	6.62
Year 2010	0.109***	6.98	0.144***	10.09
Winter	−0.257***	−16.13	−0.254***	−15.82
Spring	0.048***	3.00	0.051***	3.18
Summer	0.198***	10.84	0.203***	10.78
Migration dummy*Summer	0.106***	10.03	0.103***	9.66
Migration dummy*Fall	0.119***	6.71	0.121***	6.85
Means for random parameters:				
Constant	1.871***	20.95	2.701***	31.08
Vehicles share	0.029***	22.00	0.026***	20.55
Rain amount	−0.001***	−7.33	−0.001***	−6.53
Share of the red light rule violation	0.002*	1.74	0.002***	2.83
Share of the speed limit rule violation	0.004***	5.75	0.004***	11.00
Scale parameters for distribution of random parameters:				
Constant	0.485***	71.11	0.650***	28.55
Vehicles share	0.076***	84.25	0.001	1.53
Rain amount	0.001***	13.56	0.000***	5.58
Share of the red light rule violation	0.011***	22.02	0.009***	14.30
Share of the speed limit rule violation	0.040***	33.49	0.003***	24.65
Dispersion parameter (α)	69.432***	22.41	67.814***	22.57
Other statistical information:				
Log-likelihood values at convergence (LL_c)		−11,247.824		−11,217.075
Restricted log-likelihood value (LL_r)		−1,179,724.853		−1,179,724.853
AIC		22,563.648		22,522.150
BIC		22,766.899		22,785.181
Pseudo-R^2		0.304		0.306
Predicted numbers of accidents		94.180		94.161

Note that the degrees of freedom for LL_r is 5 and 15 for uncorrelated and correlated random parameters models since the model compare them to the conventional negative binomial model with pooled data.
* $p < 0.10$;** $p < 0.05$;*** $p < 0.01$.

Fixed

Random

23 variables

4 variables

Models for better capturing unobserved heterogeneity

Random-Parameters NB Models (RPNB)

Parameters	NB		RPNB		NB-L		RPNB-L	
	Value	Std. Dev.	Value	Std. Dev.	Value	Std. Dev.	Value	Std. Dev.
Parameter Mean								
Intercept	-4.449	0.067	-5.486	0.035	-3.947	0.162	-4.443	0.206
Log(ADT)	0.689	0.133	0.816	31.750	0.651	0.145	0.717	0.231
Friction	-0.027	0.011	-0.029	0.133	-0.027	0.012	-0.032	0.015
Pavement	0.422	0.189	0.588	0.012	0.445	0.210	0.605	0.281
Median Width	-0.005	0.002	-0.012	0.240	-0.006	0.002	-0.012	0.004
Barrier	-3.031	0.308	-6.614	0.003	-3.282	0.338	-6.152	0.898
Rumble	-0.405	0.186	-0.288	0.437	-0.404	0.207	-0.329	0.260
$\alpha = 1/\phi$	0.950	0.122	0.137	0.035	0.239	0.083	0.128	0.028
θ					1.464	0.180	1.414	0.173
Std. Deviation of Random Parameters								
Log(ADT)			0.302	0.172			0.232	0.137
Friction			0.057	0.011			0.056	0.011
Pavement			0.326	0.216			0.291	0.200
Median Width			0.028	0.003			0.028	0.003
Barrier			2.390	0.399			1.925	0.709
Rumble			0.379	0.242			0.310	0.183
Model Performance								
Dbar	1891.93		1481.09		1585.93		1422.70	
Dhat	1883.01		1296.86		1469.51		1276.00	
pD	8.92		184.22		116.41		146.30	
DIC	1900.84		1665.31†		1702.34		1569.00	
MAD[5]	6.92		6.90		6.88		6.71	

Note: † With the MLE RPNB, only three variables (logarithm of ADT, presence of median barrier and interior rumble strips) were found to be random. This increased the Deviance Information Criterion or DIC to 1736.

Models for better capturing unobserved heterogeneity

Random-Parameters NB Models (RPNB)

Parameters	NB		RPNB		NB-L		RPNB-L	
	Value	Std. Dev.	Value	Std. Dev.	Value	Std. Dev.	Value	Std. Dev.
Parameter Mean								
Intercept	-4.449	0.067	-5.486	0.035	-3.947	0.162	-4.443	0.206
Log(ADT)	0.689	0.133	0.816	31.750	0.651	0.145	0.717	0.231
Friction	-0.027	0.011	-0.029	0.133	-0.027	0.012	-0.032	0.015
Pavement	0.422	0.189	0.588	0.012	0.445	0.210	0.605	0.281
Median Width	-0.005	0.002	-0.012	0.240	-0.006	0.002	-0.012	0.004
Barrier	-3.031	0.308	-6.614	0.003	-3.282	0.338	-6.152	0.898
Rumble	-0.405	0.186	-0.288	0.437	-0.404	0.207	-0.329	0.260
$\alpha = 1/\phi$	0.950	0.122	0.137	0.035	0.239	0.083	0.128	0.028
θ					1.464	0.180	1.414	0.173
Std. Deviation of Random Parameters								
Log(ADT)			0.302	0.172			0.232	0.137
Friction			0.057	0.011			0.056	0.011
Pavement			0.326	0.216			0.291	0.200
Median Width			0.028	0.003			0.028	0.003
Barrier			2.390	0.399			1.925	0.709
Rumble			0.379	0.242			0.310	0.183
Model Performance								
Dbar	1891.93		1481.09		1585.93		1422.70	
Dhat	1883.01		1296.86		1469.51		1276.00	
pD	8.92		184.22		116.41		146.30	
DIC	1900.84		1665.31†		1702.34		1569.00	
MAD[5]	6.92		6.90		6.88		6.71	

Note: † With the MLE RPNB, only three variables (logarithm of ADT, presence of median barrier and interior rumble strips) were found to be random. This increased the Deviance Information Criterion or DIC to 1736.

Multi-distribution models

Negative Binomial-Lindley Model (NB-L)

The PMF of the NB-L regression for y_i is

$$P(y \mid \mu, \phi, \theta) = \int \mathrm{NB}(y \mid \phi, \varepsilon\mu)\mathrm{Lindley}(\varepsilon \mid \theta)\, d\varepsilon$$

The mean and variance are given by

$$E(y) = \mu \times E(\varepsilon) = \exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)\frac{\theta+2}{\theta(\theta+1)}$$

$$Var(y) = \mu \times \frac{\theta+2}{\theta(\theta+1)} + \mu^2 \times \frac{2(\theta+3)}{\theta^2(\theta+1)} \times \frac{(1+\phi)}{\phi} - \left(\mu \times \frac{\theta+2}{\theta(\theta+1)}\right)^2$$

The parameterization described above can be modified and framed as a hierarchical model (Bayesian). See Geedipally et al. (2012).

Statistical Models For Crash Data

Selection between NB and NB-L

If the skewness is greater than 1.92, use the NB-L:

Summary Statistics	Michigan Dataset
Mean	0.68
Variance	3.15
Standard Deviation (Sd.)	1.77
Variance-to-Mean-Ratio (VMR)	4.62
Coefficient-of-Variation (CV)	2.60
Skewness (skew)	7.76
Kurtosis (K)	123.59
Percentage-of-Zeros (Z)	69.6%
10-th Quantile	0
20-th Quantile	0
30-th Quantile	0
40-th Quantile	0
50-th Quantile (Median)	0
60-th Quantile	0
70-th Quantile	1
80-th Quantile	1
90-th Quantile	2
10-th Inter-Quantile	1
20-th Inter-Quantile	1
30-th Inter-Quantile	1
40-th Inter-Quantile	2
Range	61

Semi- and nonparametric models

Generalized additive models (GAMs)

The relationship between the mean and the parameters can be defined as follows

$$\mu_i = \exp\left(\beta_0 + \sum_{j=1}^{p} f_j(x_{ij})\right)$$

where β_0 is the intercept of the model and $f_j(x_{ij})$ is the smooth function (e.g., P-splines, kernel cubic regression splines, smoothers, and thin-plate regression splines). The generalized additive models can also include a combination of fixed, nonlinear functions or a combination of nonlinear functions:

$$\mu_i = \exp\left(\beta_0 + \sum_{j=1}^{k} x_{ij} + f_{k12}\left(x_{i(k+1)}, x_{i(k+2)}\right) + \sum_{j=k+3}^{p} f_j(x_{ij})\right)$$

Semi- and nonparametric models

Generalized additive models (GAMs)

$$\mu_i = \exp(\beta_0 + f_1(x_{i1}) + f_2(x_{i2}))$$

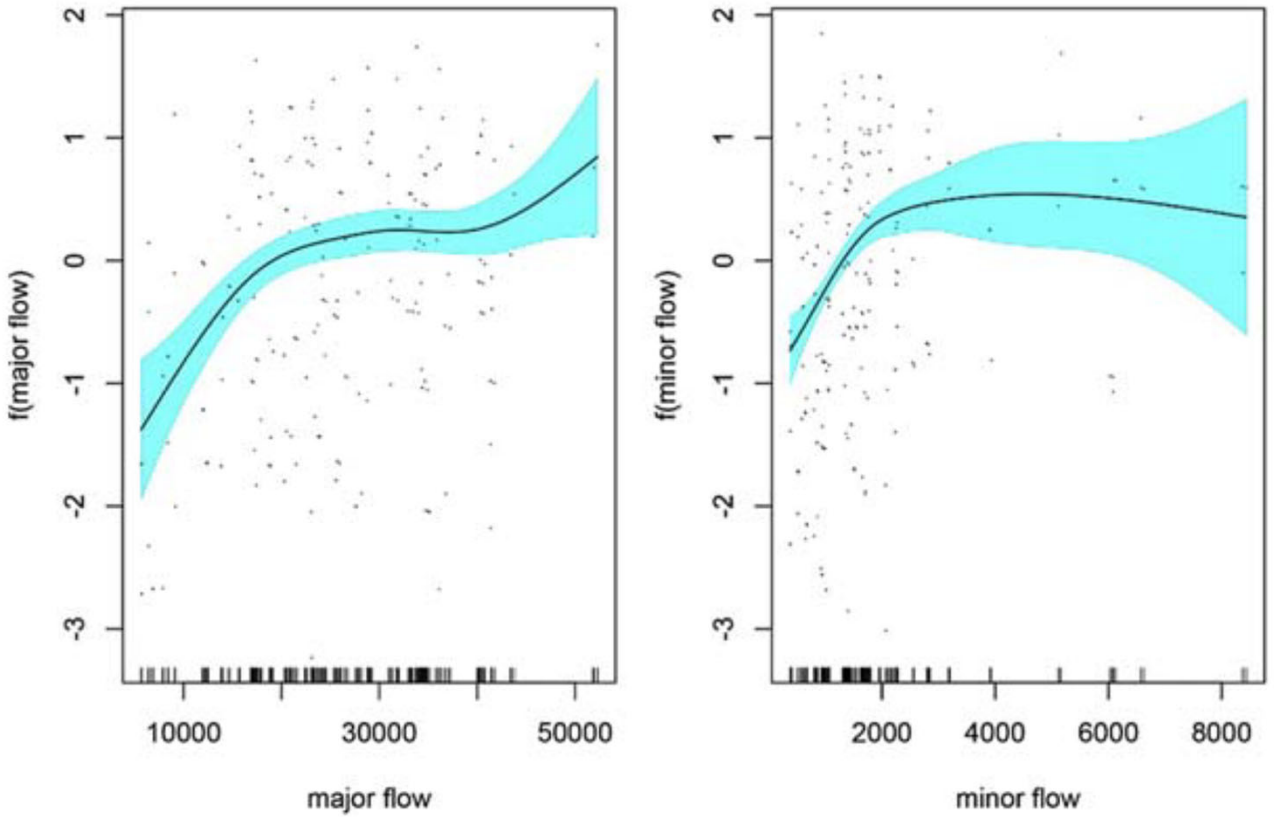


FIGURE 3.4 Relationship between the number of crashes and entering flows (Xie and Zhang, 2008).

Semi- and nonparametric models

The seminonparametric (SNP) Poisson model

For this model, the unobserved heterogeneity captured by the model's error (ε) follows a K-not polynomial function.

$$f(\varepsilon) = \frac{\left(\sum_{m=0}^{k} a_m \varepsilon^m\right)^2 \phi(\varepsilon)}{\int_{-\infty}^{+\infty} \left(\sum_{m=0}^{k} a_m \varepsilon^m\right)^2 \phi(\varepsilon) d\varepsilon}$$

where "K" refers to the length of the polynomial, "m" is an indicator increasing from 0 to "K", am is a constant coefficient, and φ(ε) represents the PDF of the standard normal distribution.

Semi- and nonparametric models

The seminonparametric (SNP) Poisson model

By mixing the SNP with the Poisson distribution, the SNP-Poisson model can be defined as follows:

$$P(y_i|\mathbf{x}_i) = \int_{-\infty}^{+\infty} P(y_i|\varepsilon_i) f(\varepsilon_i) d\varepsilon_i$$

$$P(y_i|\mathbf{x}_i) = \int_{-\infty}^{+\infty} \left\{ \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i} \times \frac{\left(\sum_{m=0}^{k} a_m \varepsilon_i^m \right)^2 \phi(\varepsilon_i)}{\sum_{m=0}^{K} \sum_{n=0}^{K} a_m a_n I(m+n)} \right\}.$$

Semi- and nonparametric models

The seminonparametric (SNP) Poisson model

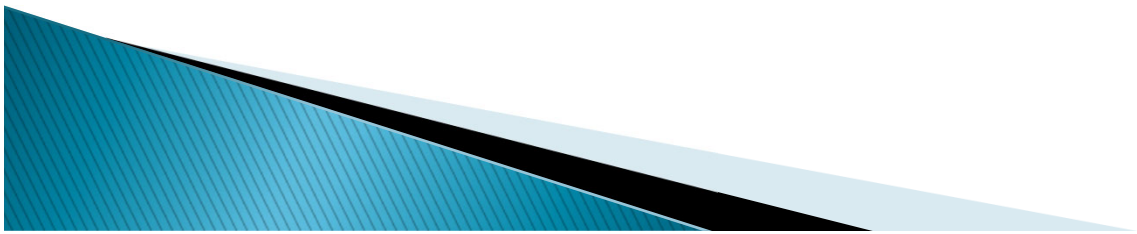
As the unconditional probability function of the previous equation does not have a closed form, the numerical method of the Gausse- Hermite quadrature needs to be applied to approximate the unconditional probability:

$$P(y_i|\mathbf{x}_i) = \sum_{j=1}^{J} \left\{ w_j \left\{ \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i} \right\} \times \left[\frac{\left(\sum_{m=0}^{k} a_m \varepsilon_i^m\right)^2 \phi(\varepsilon_i)}{\sum_{m=0}^{K}\sum_{n=0}^{K} a_m a_n I(m+n)} \right] \right\}.$$

Semi- and nonparametric models

Dirichlet process models

The DP is a stochastic process that is usually used as a prior in Bayesian nonparametric (or semiparametric) modeling. In this regard, Escobar and West (1998) defined the DP as a random probability measure over the space of all probability measures. In that sense, the DP is considered as a distribution over all possible distributions; that is, each draw from the DP is itself a distribution, which may not be same as for the previous draw.



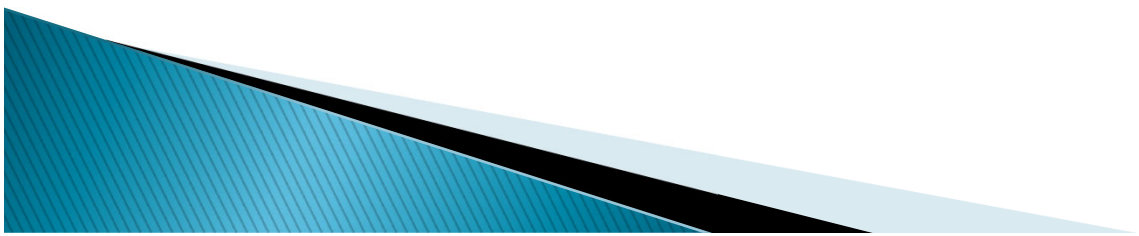
Semi- and nonparametric models

Dirichlet process models

Models that employ the Dirichlet process (DP), widely used in the Bayesian literature, can technically be classified as either nonparametric or semiparametric models depending on the modeling framework.

For semiparametric models, as applied with safety data, the count data still follow a Poisson distribution, but the mean or the error term is assumed to follow a Dirichlet distribution or process.

As opposed to the Poisson or NB mixtures, in which two parametric distributions are mixed together, the DP is characterized by an infinite mixture of distributions, where the number of unique components or distributions and the component characteristics themselves can be learned from the data.



Semi- and nonparametric models

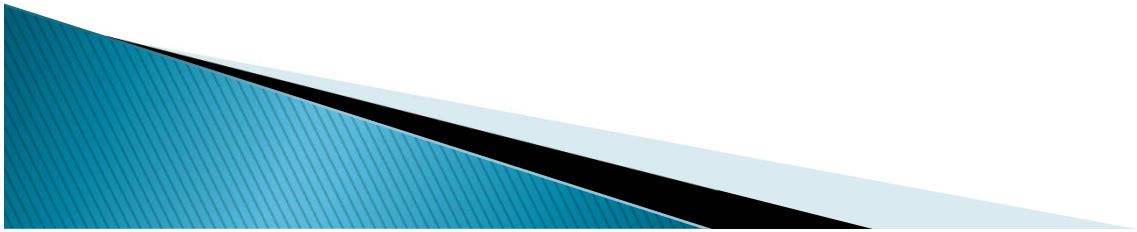
Dirichlet process models

In the safety literature, two semiparametric models have been proposed, one for the Poisson, called the Poisson-Dirichlet Process (P-DP), and one for the multi-parameter NB model, called the NB-DP. For the P-DP, we have this formulation:

$$P(y|\mu,\tau,F(.|\theta)) = \int \text{Poisson}(y|\nu\mu)dF(\nu|\text{DP}(\tau,F(.|\theta))).$$

For the NB-DP, we have the following:

$$P(y|\mu,\phi,\tau,F(.|\theta)) = \int \text{NB}(y|\nu\mu,\phi)dF(\nu|\text{DP}(\tau,F(.|\theta))).$$



Semi- and nonparametric models

Dirichlet process models

As stated in the previous section, the distribution of $\mathrm{DP}(\tau, F(.|\theta))$ can be approximated by its truncated construction $\mathrm{TDP}(\tau, M, F(.|\theta))$. Consequently, the P-TDP and NB-TDP model can be seen as a hierarchical Bayesian model described as follows:

$$P(y_i|\nu_i\mu_i) = \mathrm{Poisson}(\nu_i\mu_i)$$

$$P(y_i|\nu_i\mu_i, \phi_i) = \mathrm{NB}(\nu_i\mu_i, \phi)$$

$$\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$$

$$\gamma_k|\tau \sim \mathrm{Beta}(1, \tau), \quad k = 1, 2, \ldots, M$$

$$\psi_k|\theta \sim F_0(.|\theta), \quad k = 1, 2, \ldots, M$$

$$p_k = \gamma_k \prod_{k'<k} (1 - \gamma_{k'}), \quad k = 1, 2, \ldots, M$$

$$\nu_i \sim F()$$

$$F() \sim \mathrm{TDP}(\tau, F_0(.|\theta)) \equiv \sum_k p_k \delta_{\psi_k}$$

Semi- and nonparametric models

Dirichlet process models

TABLE 3.1 Modeling Results for the Indiana Data (Shirazi et al., 2017).

Variables	NB			NB-L			NB-TDP		
	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)
Intercept (β_0)	−4.779	0.979	0.0000	−3.739	1.115	0.0008	−7.547	0.1227	0.0000
Ln(ADT) (β_1)	0.722	0.091	0.0000	0.630	0.106	0.0000	0.983	0.117	0.0000
Friction (β_2)	−0.02774	0.008	0.0006	−0.02746	0.011	0.1300	−0.01999	0.008	0.0126
Pavement (β_3)	0.4613	0.135	0.0005	0.4327	0.217	0.0468	0.3942	0.152	0.0100
MW (β_4)	−0.00497	0.001	0.0000	−0.00616	0.002	0.0021	−0.00468	0.002	0.0195
Barrier (β_5)	−3.195	0.234	0.0000	−3.238	0.326	0.0000	−8.035	1.225	0.0000
Rumble (β_6)	−0.4047	0.131	0.0021	−0.3976	0.213	0.0609	−0.3780	0.150	0.0134
$\alpha = 1/\phi$	0.934	0.118	0.0000	0.238	0.083	0.0074	0.301	0.085	0.0042
DIC[a]		1900			1701			1638	
MAD[b]		6.91			6.89			6.63	
MSPE[c]		206.79			195.54			194.5	

[a] *Deviance information criterion.*
[b] *Mean absolute deviance.*
[c] *Mean squared predictive error.*

Semi- and nonparametric models

Dirichlet process models

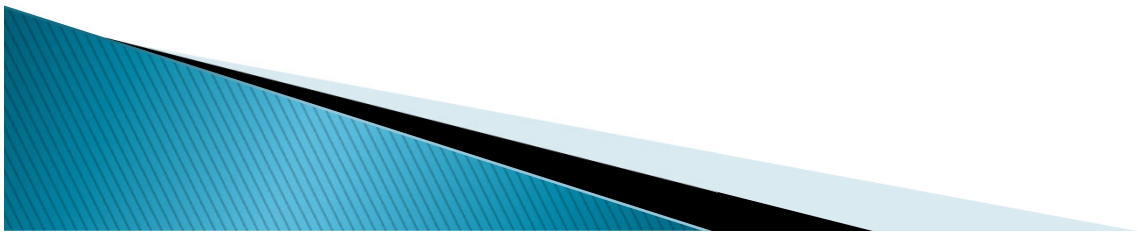
Site	1	2	3	4	5	6	7	8	9	10
1	1.0	0.6	0.6	0.6	0.6	0.2	0.6	0.6	0.1	0.1
2	0.6	1.0	0.6	0.6	0.6	0.2	0.6	0.6	0.1	0.1
3	0.6	0.6	1.0	0.6	0.6	0.2	0.6	0.6	0.1	0.1
4	0.6	0.6	0.6	1.0	0.6	0.2	0.6	0.6	0.1	0.1
5	0.6	0.6	0.6	0.6	1.0	0.2	0.6	0.6	0.1	0.1
6	0.2	0.2	0.2	0.2	0.2	1.0	0.2	0.2	0.6	0.6
7	0.6	0.6	0.6	0.6	0.6	0.2	1.0	0.6	0.1	0.1
8	0.6	0.6	0.6	0.6	0.6	0.2	0.6	1.0	0.1	0.1
9	0.1	0.1	0.1	0.1	0.1	0.6	0.1	0.1	1.0	0.6
10	0.1	0.1	0.1	0.1	0.1	0.6	0.1	0.1	0.6	1.0

FIGURE 3.5 The heatmap representation of the partitioning matrix for the top 10 sites with the highest ADT values in the Indiana dataset (Shirazi et al., 2018).

Semi- and nonparametric models

Nonparametric models

- Nonparametric models have been used relatively often in highway safety.
- Popular ones: multilayer perceptron (MLP) neural network, convolutional neural networks, Bayesian neural networks (BNN), and support vector machine (SVM).
- Usually good for predicting crashes, but can easily over-fit the data.
- Caution: They work as black boxes; need to use tools for examining sensitivity.
- Chapter 12 covers these models in greater details.



Model Selection

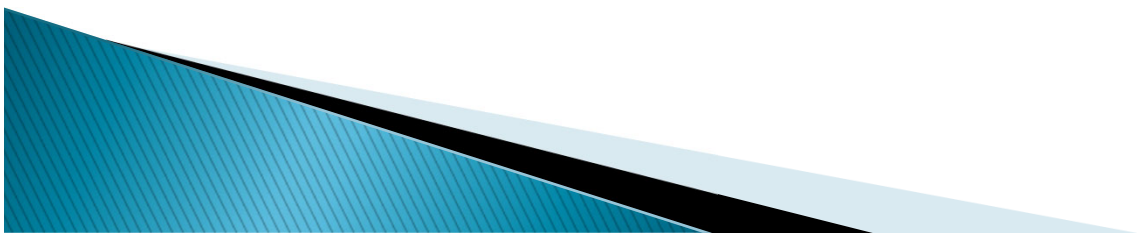
- In this lecture, we have presented different types of models that vary from the most basic to very complex.
- Positive and negative attributes have been provided for most of these models, along with a description explaining when the model could be suitable given the known and unknown characteristics of the data.
- In the highway safety literature, a lot of work has been devoted to the development and application of statistical models (in fact, this is the majority of the work produced in highway safety according to Zou and Vu, 2019).

Model Selection

- A common theme of a new study is that the "new" proposed model is claimed to be better than other previously published or widely applied models because it fits the data better.
- In other words, the "new" model reduces the unobserved heterogeneity more than the previous model.
- The model also needs to adequately capture the data generating process of the dataset under study. Miaou and Lord (2003) refer to this subject as the "goodness-of-logic," which consists of making sure the model properly characterizes the analyzed data and the model is methodologically sound.
  - Example Zero-Inflated/Hurdle Models (see textbook)

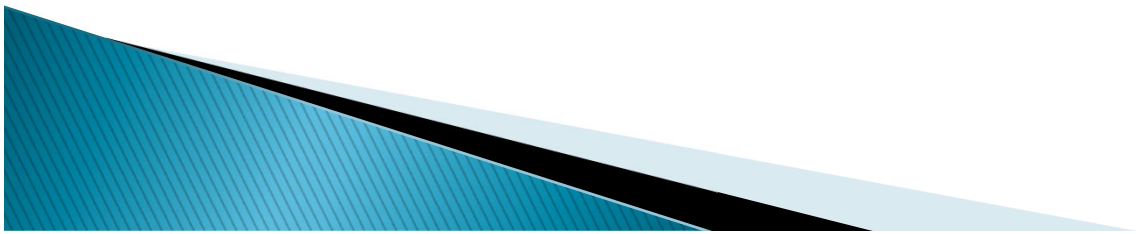
Model Selection

- Along the same line, using a very complex model does not necessarily mean that the model is better, even if the "fit" is superior.
- The model could be overly complex given the study objectives or the gains they provide compared to traditional models are considered marginal.



Model Selection

- Complex models often do not have the opportunity to be fully validated using theoretical principles, simulation, or a wide range of datasets, especially as many have been recently been introduced in the safety literature.
- It is not uncommon to see new models that are later found to suffer from important methodological limitations (e.g., zero-inflated models).
- Depending on the parameterization and the estimation method, the model could also take a very long time to provide results. For example, some models estimated using the Bayesian method can sometimes take several days or hours for the MCMC posterior estimates to converge.
  - A pragmatic or better approach would be to use the MLE method when it can be used based on the study objectives and characteristics of the data, but to use the Bayesian estimation method when the MLE cannot be used because of the complexity of the model.



Model Selection

- Final thoughts :
  - *Although the NB model can become unreliable under particular conditions, this model, which is characterized by solid theoretical foundations, has had the opportunity to be analyzed, tweaked and used by researchers and practitioners across the globe for several decades and is considered more than adequate for most applications.*
  - *As Dr. George Box famously said "all models are wrong but some are useful" (Box, 1979) and "the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity" (Box, 1976), meaning that more complex models are not necessarily better.*